UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

| APPLICATION NO. | FILING DATE | FIRST NAMED INVENTOR | ATTORNEY DOCKET NO. | CONFIRMATION NO. |
|---|---|---|---|---|
| 09/944,332 | 08/30/2001 | Liqin Shen | JP920000191US1 (590.079) | 1865 |

35195 7590 10/17/2007

FERENCE & ASSOCIATES LLC
409 BROAD STREET
PITTSBURGH, PA 15143

| EXAMINER | |
|---|---|
| HAN, QI | |

| ART UNIT | PAPER NUMBER |
|---|---|
| 2626 | |

| MAIL DATE | DELIVERY MODE |
|---|---|
| 10/17/2007 | PAPER |

**Please find below and/or attached an Office communication concerning this application or proceeding.**

The time period for reply, if any, is set in the attached communication.

| Office Action Summary | Application No. | Applicant(s) | |
|---|---|---|---|
| | 09/944,332 | SHEN ET AL. | |
| | Examiner | Art Unit | |
| | Qi Han | 2626 | |

*-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address --*

**Period for Reply**

A SHORTENED STATUTORY PERIOD FOR REPLY IS SET TO EXPIRE 3 MONTH(S) OR THIRTY (30) DAYS, WHICHEVER IS LONGER, FROM THE MAILING DATE OF THIS COMMUNICATION.

- Extensions of time may be available under the provisions of 37 CFR 1.136(a). In no event, however, may a reply be timely filed after SIX (6) MONTHS from the mailing date of this communication.
- If NO period for reply is specified above, the maximum statutory period will apply and will expire SIX (6) MONTHS from the mailing date of this communication.
- Failure to reply within the set or extended period for reply will, by statute, cause the application to become ABANDONED (35 U.S.C. § 133). Any reply received by the Office later than three months after the mailing date of this communication, even if timely filed, may reduce any earned patent term adjustment. See 37 CFR 1.704(b).

**Status**

1)☒ Responsive to communication(s) filed on *24 July 2007*.

2a)☐ This action is **FINAL**. 2b)☒ This action is non-final.

3)☐ Since this application is in condition for allowance except for formal matters, prosecution as to the merits is closed in accordance with the practice under *Ex parte Quayle*, 1935 C.D. 11, 453 O.G. 213.

**Disposition of Claims**

4)☒ Claim(s) *1-19* is/are pending in the application.

4a) Of the above claim(s) _____ is/are withdrawn from consideration.

5)☐ Claim(s) _____ is/are allowed.

6)☒ Claim(s) *1-19* is/are rejected.

7)☐ Claim(s) _____ is/are objected to.

8)☐ Claim(s) _____ are subject to restriction and/or election requirement.

**Application Papers**

9)☐ The specification is objected to by the Examiner.

10)☐ The drawing(s) filed on _____ is/are: a)☐ accepted or b)☐ objected to by the Examiner.

Applicant may not request that any objection to the drawing(s) be held in abeyance. See 37 CFR 1.85(a).

Replacement drawing sheet(s) including the correction is required if the drawing(s) is objected to. See 37 CFR 1.121(d).

11)☐ The oath or declaration is objected to by the Examiner. Note the attached Office Action or form PTO-152.

**Priority under 35 U.S.C. § 119**

12)☐ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).

a)☐ All b)☐ Some * c)☐ None of:

1.☐ Certified copies of the priority documents have been received.

2.☐ Certified copies of the priority documents have been received in Application No. _____.

3.☐ Copies of the certified copies of the priority documents have been received in this National Stage application from the International Bureau (PCT Rule 17.2(a)).

* See the attached detailed Office action for a list of the certified copies not received.

**Attachment(s)**

1) ☐ Notice of References Cited (PTO-892)
2) ☐ Notice of Draftsperson's Patent Drawing Review (PTO-948)
3) ☐ Information Disclosure Statement(s) (PTO/SB/08) Paper No(s)/Mail Date _____.
4) ☐ Interview Summary (PTO-413) Paper No(s)/Mail Date. _____.
5) ☐ Notice of Informal Patent Application
6) ☐ Other: _____.

## DETAILED ACTION

1. The text of those sections of Title 35, U.S. Code not included in this action can be found in a prior Office action.

### *Continued Examination Under 37 CFR 1.114*

2. A request for continued examination under 37 CFR 1.114, including the fee set forth in 37 CFR 1.17(e), was filed in this application after final rejection. Since this application is eligible for continued examination under 37 CFR 1.114, and the fee set forth in 37 CFR 1.17(e) has been timely paid, the finality of the previous Office action has been withdrawn pursuant to 37 CFR 1.114.

### *Response to Amendment*

3. This communication is responsive to the applicant's amendment and RCE examination both filed on 07/24/2007. The applicant(s) amended claims 1, 10 and 19 (see the amendment: pages 2-6).

It is noted that applicant and examiner discussed the issue regarding the 112 rejection during in the interviews on 10/3/2007 and 10/05/2007. The examiner withdraws the claim rejection regarding the previous amended limitation "wherein the segmenting and the splitting is **not dependent upon word boundaries**", under 35 USC 112 1st and 2nd. As a record of the prosecution of this application, the examiner should point out that, this withdrawal is only based on the original specification disclosure, which states "for some language, however, such as Chinese and Japans, there is no word boundary in written language, and words are not well

defined..." (specification: bridge paragraph between pages 1-2). The examiner disagrees with the applicant's arguments that the above limitation "can also be utilized for languages in which boundaries exit", such as "English language" (see Remark: page 8, paragraph 2), because the original specification has no any specific description about the argument and/or one of ordinary skill in the art could not solve the extremely high-volume computation problem brought from the argued exemplary method for English language.

It is also noted that applicant has provided a proposed amendment, in which the third element of the independent claims includes an newly added limitation "wherein the new words are words not contained in a base vocabulary". Even though the claim provides further limitation and clarification, it is noted that Yang discloses solving "out of vocabulary (OOV)" problem for language model and lexicon, which can be properly reads on the claimed limitation. Therefore, the proposed claim amendment cannot overcome the prior art rejection based on the combined reference teachings. Further, since the proposed amendment has not been formally entered/scanned in the PTO working database and the proposed claim is not in condition of allowance, the examine cannot consider to enter it by the examiner's amendment. Accordingly, at this point, this office action will be based on the amendment filed on 07/24/2007 (which is the latest version of the amendment in the PTO IF working database), hereinafter (see below).

***Response to Arguments***

4. Applicant's arguments filed on 07/24/2007 with respect to the claim rejection under 35 USC 103, have been fully considered but are moot in view of the new ground(s) of rejection, since the amended claims introduce new issue and/or change the scope of the claims. It is noted

that the previous cited reference are still applicable to the amended claims for the prior art rejection. The response to the applicant's arguments is directed to the corresponding claim rejection (see detail below).

5. Further, in response to applicant's arguments against the references individually (see Remarks: page 9, paragraph 2 to page 10, paragraph 3), one cannot show nonobviousness by attacking references individually where the rejections are based on combinations of references. See *In re Keller*, 642 F.2d 413, 208 USPQ 871 (CCPA 1981); *In re Merck & Co.*, 800 F.2d 1091, 231 USPQ 375 (Fed. Cir. 1986).

6. Furthermore, in response to applicant's argument (Remarks: page 10, paragraph 4), that there is no suggestion to combine the references, the examiner recognizes that obviousness can only be established by combining or modifying the teachings of the prior art to produce the claimed invention where there is some teaching, suggestion, or motivation to do so found either in the references themselves or in the knowledge generally available to one of ordinary skill in the art. See *In re Fine*, 837 F.2d 1071, 5 USPQ2d 1596 (Fed. Cir. 1988)and *In re Jones*, 958 F.2d 347, 21 USPQ2d 1941 (Fed. Cir. 1992). In this case, the obviousness is based the same reason and/or scope in the previous office action (see rejection for claim 1), because it can properly cover the amended claims and the applicant's argument only comprised assertion(s) without any detailed analysis (see Remarks: page 10, paragraph 4).

***Claim Rejections - 35 USC § 103***

7. Claims 1-3, 6-12 and 15-19 are rejected under 35 U.S.C. 103(a) as being unpatentable over Wang et al. (US 6,904,402 B1) hereinafter referenced as Wang, in view of Razin et al. (US 6,098,034) hereinafter referenced as Razin and Yang et al. ("statistics-based segment pattern lexicon—a new direction for Chinese language modeling", 0-7803-4428-6/98, IEEE, pp 169-172) hereinafter referenced as Yang.

As per **claim 1**, Wang discloses system and iterative method for lexicon, segmentation and language model joint optimization (title), comprising:

"segmenting a cleaned corpus [in a domain] to form a segmented corpus", (Fig. 5 and col. 9, lines 36-44, 'segmentation', 'the received corpus is built', 'pre-processed to remove some obvious illogical words (so as to provide cleaned corpus)');

"splitting the segmented corpus to form sub strings, and counting the occurrences of each sub strings appearing in the corpus" (col. 1, lines 45-60, 'a textual corpus is dissected (interpreted as split) into a plurality of items (sub strings)' and 'counts the number of occurrences of a particular item (word, character, etc.)'); and

Even though Wang further suggests that 'the items of the corpus' having low occurrence frequency 'may be pruned' (col. 7, lines 27-29) and 'counting the occurrence of strings of characters' (corresponding to new words and is capable of outputting), Wang does not expressly disclose "**filtering** out false candidates to output new words" and segmenting the corpus "in a **domain**" . However, this feature is well known in the art as evidenced by Razin who, in the same field of endeavor, discloses method for standardizing phrasing in a document (title), comprising 'filtering the preliminary list of extracted phrases (candidates) to create (output) a

final list of extracted phrases (corresponding to or necessarily including **new words**)' (Fig. 2 and col. 30, lines 55-56), and 'domain' that 'is defined as a particular field of discourse having its own specialized terminology and type of document' (col. 7, lines 19-21) and 'identifying structural elements and their types in the document' through 'use of domain-dependent method' (col. 8, lines 1-7). Razin further discloses using 'suffix tree' and 'phase identification by establishing word sequences that satisfy the criteria for length and recurrence in the document', wherein 'each node of the tree is associated with a record of the number of occurrences of the word sequence' (col. 2, lines 3-14), which further supports the rejection stated above and the combination of the prior art teachings. Therefore, it would have been obvious to one of ordinary skill in the art at the time the invention was made to modify Wang by providing filtering a set of extracted phrases and creating (output) final phrases list (including new words) and using the documents in a domain, as taught by Razin, for the purpose (motivation) of obtaining extracted words constituting significant user phrases (or new words) and/or providing specialized terminology in a particular field (Razin: col., 2, lines 46-47; col. 7, lines 19-20).

It is noted that Wang in view Razin does not **expressly** disclose "the segmenting and the splitting is not dependent upon word boundaries" and "wherein new words are determined based upon the domain of the cleaned corpus" . However, the feature is well known in the art as evidenced by Yang who, in the same field of endeavor, discloses 'statistics-based segment pattern lexicon—a new direction for Chinese language modeling' (title), teaches that since 'there are no "blanks" in Chinese sentences serving as word boundaries, ...the "word" in Chinese are actually not well defended' (abstract), so that the elements in the lexicon called 'segment pattern of characters' 'should be extracted form the training corpus (corresponding to clean corpus) from

the training corpus by statistical approach (that is not dependent upon word boundaries)' (page 169, right col., paragraph 3). Further, Yang teaches that 'a new lexicon (including new words) is certainly needed' and 'the element in this new lexicon can be either words, or phrases... commonly accepted templates, etc., many of which are "out of vocabulary (OOV)" (i.e. new words) for most conventional lexicons...' (page 169, right col., paragraph 3) and 'segment pattern extraction approach' using 'prefix and suffix trees' for 'all character strings occurring in the training corpus' (page 170, left col., paragraph 2), which further supports the rejection stated above and the combination of the prior art teachings. Therefore, it would have been obvious to one of ordinary skill in the art at the time the invention was made to recognize that the OOV (new words) extracted from the training corpus (cleaned corpus) are necessarily determined based on a domain of the corpus, such as 'the domain of legal documents (corpus)' as disclosed by Rezin (Rezin: col. 8, line 3), and to modify Wang in view of Razin by providing segment pattern extraction for the character-based language models, such as Chinese language, for the new lexicon elements (new words) extracted from the training corpus by a statistical approach, as taught by Yang, for the purpose (motivation) of minimizing the overall perplexity for the segmentation and/or solving OOV problem of processing character-based language (Yang: abstract and page 169, right col., paragraph 2).

Moreover, in another view of disclosure of Wang and Razin, Wang further discloses a system and method 'for lexicon, segmentation and language model joint optimization' (col. 2, lines 43-56), and teaches that 'a language model can take any sequence of items (**words**, **charters**, letters, etc.) and estimate the probability of the sequence' (col. 1, line 35-41) and providing 'a dynamic segmentation function 216 to segment items (**characters** or letters, for

example) into strings (e.g., words)', which suggests that the system/method has capability to perform a character-based segmentation. Wang further disclose 'the prefix tree may be built using the entire corpus, or alternatively, using a subset entire corpus (referred to as a training corpus)' for the lexicon generation (col. 10, lines 30-67) and 'to optimize a statistical language model from the received corpus (or training set)' using 'the segmented corpus (cleaned corpus)' (col. 11, lines 6-18). Therefore, it would have been obvious to one of ordinary skill in the art at the time the invention was made to recognize that the most popular eastern languages, such as Chinese or Japanese, are character-based languages and have no blank or space served as word boundaries in the written form and lexicon generation from the training corpus (cleaned corpus) would be based on a domain of the corresponding training corpus, so that the combined system/method from Wang in view of Razin can perform a segmentation for those character-based languages and generation of lexicon from the training corpus in domain (such as in 'the domain of legal documents (corpus)' as disclosed by Rezin: col. 8, line 3), as Wang suggested, for the purpose (motivation) of improving language model performance and/or providing capability of segmenting items (including characters) into words for a textural corpus (Wang: col. 2, line 50-51 and col. 1, lines 52-59). This means that only Wang in view of Razin, can provides sufficient basis for the rejection, based on broad interpretation of the claim.

As per **claim 2** (depending on claim 1), Wang in view of Razin and Yang further discloses "using punctuations, Arabic digits and alphabetic strings, or new words patterns to split the cleaned corpus", (Razin, col. 21, lines 10, 'punctuation'; col. 4, lines 26, 'the usage of stop list');

As per **claim 3** (depending on claim 1), Wang in view of Razin and Yang further discloses "using common vocabulary to segment the cleaned corpus", (Razin: col. 5, lines 36-45, 'the dictionary of standard phrases (common vocabulary)').

As per **claim 6** (depending on claim 1), Wang in view of Razin and Yang further discloses:

"filtering out functional words" (Razin: col. 4, lines 35-38, 'stop list', 'semantically insignificant words (e.g., "and then about the") (interpreted as functional words)', which suggests that these words can be filtered out);

"filtering out those sub strings which almost always appear along with a longer sub strings" (Razin: col. 9, lines 52, 'eliminates from the phrase list otherwise-significant phrases that are nested within other significant phrases... removes from the final phrase list minimal content words dangling at the beginning or end of preliminary user-specific phrases', which reads on the claim); and

"filtering out those sub strings for which the occurrence is less than a predetermined threshold", (Razin: col. 2, lines 10-13, 'each node of tree is associated with a record of the number of occurrence of the word sequence at that node, where the number of occurrence exceeds the required threshold', which reads on the claimed limitation).

As per **claim 7** (depending on claim 1), Wang in view of Razin and Yang further discloses "using pre-recognized functional words as segment boundary patterns", (Razin: col. 4, lines 35-38, 'stop list', 'semantically insignificant words (e.g., "and then about the") (interpreted as functional words)').

As per **claim 8** (depending on claim 3), the rejection is based on the same reason described for claim 7 because the claim recites the same or similar limitation(s) as claim 7.

As per **claim 9** (depending on claim 3), the rejection is based on the same reason described for claim 6 because the claim recites the same or similar limitation(s) as claim 6.

As per **claims 10-12 and 15-18**, they recite an automatic new word extraction system. The rejection is based on the same reason described for claims 1-3 and 6-9, respectively, because the claims recite the same or similar limitation(s) as claims 1-3 and 6-9, respectively.

As per **claim 19**, it recites a program storage device readable by machine. The rejection is based on the same reason described for claim 1, because the claim recites the same or similar limitations as claim 1.

8. Claims 4-5 and 13-14 are rejected under 35 U.S.C. 103(a) as being unpatentable over Wang in view of Razin and Yang as applied to claims 1 and 10, and further in view of Hui (IDS: "Color Set Size Problem with Applications to String Matching," Proc. of 2nd Symposium on Combinatorial Pattern Matching, 1992, pp. 230-243).

As per **claim 4** (depending on claim 1), even Wang in view of Razin and Yang further discloses using suffix tree (i.e. atomic suffix tree—AST) (Wang: col. 1, line 42; Razin: col., 2, line 3), Wang in view of Razin does not expressly disclose "using a GAST". However, the feature is well known in the art as evidenced by Hui who teaches 'the concept of suffix tree can be extended' and 'this extension is called the Generalized suffix tree (GST)( corresponding to GAST)' (Hui, page 237, first paragraph). Therefore, it would have been obvious to one of ordinary skill in the art at the time the invention was made to modify Wang in view of Razin and

Yang by specifically providing using extended suffix tree (GST or GAST), for the purpose of storing more than one input strings (Hui: page 237, first paragraph).

As per **claim 5** (depending on claim 4), Wang in view of Razin, Yang and Hui further discloses the tree "implemented by limiting length of sub strings", (Razin: col. 14, lines 34-35, 'length less than or equal to Smax').

As per **claim 13** (depending on claim 10), the rejection is based on the same reason described for claim 4 because the claim recites the same or similar limitation(s) as claim 4.

As per **claim 14** (depending on claim 10), the rejection is based on the same reason described for claim 5 because the claim recites the same or similar limitation(s) as claim 5.

***Conclusion***

9. Please address mail to be delivered by the United States Postal Service (USPS) as follows:

Mail Stop ____
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

**or faxed to:** 571-273-8300, (for formal communications intended for entry)

**Or:** 571-273-8300, (for informal or draft communications, and please label "PROPOSED" or "DRAFT")

If no Mail Stop is indicated below, the line beginning Mail Stop should be omitted from the address.

Effective January 14, 2005, except correspondence for Maintenance Fee payments, Deposit Account Replenishments (see 1.25(c)(4)), and Licensing and Review (see 37 CFR 5.1(c) and 5.2(c)), please address correspondence to be delivered by other delivery services (Federal Express (Fed Ex), UPS, DHL, Laser, Action, Purolater, etc.) as follows:

U.S. Patent and Trademark Office
Customer Window, Mail Stop _____
Randolph Building
Alexandria , VA 22314

Any inquiry concerning this communication or earlier communications from the examiner should be directed to Qi Han whose telephone numbers is (571) 272-7604. The

examiner can normally be reached on Monday through Thursday from 9:00 a.m. to 7:30 p.m. If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, Richemond Dorvil, can be reached on (571) 272-7602.

Information regarding the status of an application may be obtained from the Patent Application Information Retrieval (PAIR) system. Inquiries regarding the status of submissions relating to an application or questions on the Private PAIR system should be directed to the Electronic Business Center (EBC) at 866-217-9197 (toll-free) or 703-305-3028 between the hours of 6 a.m. and midnight Monday through Friday EST, or by e-mail at: ebc@uspto.gov. For general information about the PAIR system, see http://pair-direct.uspto.gov.

QH/qh
October 15, 2007